Week 1 Lecture 1:

From Data to Decision

Slide 2:

Our goal as analysts is to help stakeholders make decisions, and data is the raw material we use to get there. Data is not an end in and of itself, rather it is a means to a decision or solution.

What do we mean by “stakeholder”? A stakeholder is anyone within an organization with interest or concern in something.

Slide 3:

Example: Tobin’s Data Dump:

* The dean of Tobin is a data stakeholder. He wants to know why people leave the school of business. What is the “profile” of the stop-out student? Ie, What classes do these take? What is their GPA? How was their high school performance? What scholarships did they receive, etc.?

Slide 4:

So we gather some data (image of data dashboard).

Perhaps we determine that the students who leave Tobin had lower HS gpas, they struggled in the math portion of their SAT, and they are primarily commuting a long distance to class.

We present this data to the dean on a dashboard or report and call it a day. Are we done?

Slide 5:

We still don’t have a decision! This is where the not-so-straightforward part begins. We started with a problem – some students stop out of Tobin before the graduate - now have to come up with a solution to the problem. This is where statistics come in.

Statistics is a mathematical way of making inferences about a population by only observing a sample. Think “from small to big”.

Slide 6:

Solution 1

One approach is to try to get these students tutoring assistance and housing scholarships. This is expensive. Can we look at past data to see if that helped? St. John’s, and the Tobin school of business have been around for a long time, and that makes for lots of data. Surely there have been students along the way who fit the profile of the “stop out” student, but got some extra academic assistance, and/or a housing scholarship to live on campus. Did they persist in their degrees more so than their counterparts?

Slide 7:

Solution 2

Certainly our method for identifying these “at-risk” students will not be perfect. Therefore, we could end up offering expensive resources to students who would’ve retained without them, and *not* offering resources to students who needed them. These cases are called false positives and false negatives, and both carry different risks that will be unique to your business case.

Slide 8:

We aren’t going to solve this problem today, but here’s the general idea -

Data cannot make decisions, and this is not always obvious to most people – this is where statistics come in. Statistics is generalizing from specific to general, and it is a primary tool in traversing the Gulf of Data.

In solution 1, we could do what’s called a hypothesis test. We’ll cover this in chapter \_\_

In solution 2, we could create a classification model and figure out how to balance the risks of false positives vs false negatives. This will not be covered in this course, but will likely be covered in part 2 or in further analytics courses.

Slide 9:

Summing up our example - What did we do in our solutions? We looked at selected student data to try and get some idea about how the world works. In this case, we wanted to understand why students leave Tobin before graduation. Statistics is what allows us to do that. Observe the specific, infer something general.

Slide 10:

What is data?

“Values along with their context”

Simplest representation of data: the table – data organized into rows and columns.

What do the rows and columns of a table represent? Rows commonly represent “cases” and columns represent “variables”.

“Cases” in our table are customers, and the “variables” we are recording for each case are name, city, state, and zip code.

Slide 11:

Data in tabular format (or, in a table) is called *structured* data, because there is a defined schema (ie, a row defines what’s being observed, and a column defines exactly what data goes in it).

In our Customer table below, what would go in row 8? We may not know exactly what values go there, but we know there would be a customer ID, a person’s name, and their city, state, and zip code. We know this because the data clearly has structure.

For this class, and likely most of the classes in this curriculum, structured data will be the focus.

Slide 12:

Unstructured Data. Text? Audio files? Video? Images? This is called unstructured data, and is a huge component of modern data science and AI. This is currently where the cutting edge of analytics exists.

Slide 13:

Back to structured data…

Tables are traditionally stored in Relational databases.

A relational database is a way of storing tables that relate to each other. Each record in a relational database is given a “primary key” which uniquely identifies the record.

To interact with data in these databases, you must use a language called SQL (Structured Query Language). ( quick example, show related tables)

Slide 14:

Let’s look closer at our customer table. How could this go wrong? (customer table with location in column) How could this be improved? What if a customer had two addresses? Parents house vs dorm room? Apartment vs work? How could we solve this?

Sidenote: SQL is usually step 1 of solving any data-related problem in a business or organization. In this course, we’ll start at step 2…

Slide 15:

Data vs information

If structured data is defined by values and their context, what is information?

Information is the *meaning* that we draw from data.

A practical way of getting from data to information is by establishing the “Five Ws”

* who, what, when, where, why

Slide 16:

(Customer table vs Transaction table) In a table, rows are “who”, and columns are “what”. Understanding this when using tables (or spreadsheets) is extremely important.

Name the “who”s in each table. Does 6 rows in the transaction table mean 6 people ordered products?

Sidenote: sometimes tables will have timestamps (when) or location (GPS latitude/longitude), but these are not guaranteed.

Slide 17:

For the other Ws (when, where, and why), we look for metadata –

Metadata literally means “data about the data”. Often, this will tell us more detailed information about the data than we can gather by simply looking at the columns and rows. Things we might find in metadata:

* Detailed column definitions
* How the data was collected (swipe data, sensors, survey, experiment, interview, etc.)
* How it relates to other tables in a database

Many datasets will have metadata files with them, so be sure to look for them whenever they’re available. If no metadata is available, proceed with caution.

Slide 18:

Types of data (categorical and quantitative) – show a table

Data can come in many varieties, but for the purposes of this course we will focus primarily on categorical and quantitative data.

Categorical variables describe a category. Major, country, class (freshman, sophomore, etc.)

Quantitative data are numbers, and they need units.

* Sometimes these units are implied in the data, other times you will need to consult the metadata.

Find an example of a categorical variable in the tables. City and State are good examples. Find an example of a quantitative variable. Total is a good example. What are the units? Perhaps if we know more context, we can assume these are dollars. Ie, if this is a t-shirt website, this would make sense. If this is a car dealership, perhaps this represents thousands of dollars.

Are all numbers quantitative?

Slide 19:

Other types of data

Identifier variables (eg, Customer ID, Transaction ID, Student ID, etc.). Technically these would be considered categorical variables, but they are a very specific type of category

Boolean (True/False) Not covered in this course, but are very important in analytics, computer programming, and databases.

Slide 20:

Ordinal: categorical variables that have an inherent order (eg, Likert scale)

* The order is important but can’t be used for math the way quantitative variables can. For instance, 4 on Likert scale isn’t twice as much as 2.

Nominal: categorical variables that do not have inherent order (blood types, gender, martial status, etc.)

Slide 21:

Time series: describes datasets where a variable is measured over time at regular intervals (emphasis). Naturally, dates and times are always found in time series data. Time series data is modeled and analyzed using special techniques which will be introduced in part two of this course.

Cross sectional: describes datasets in which many variables are measured at a single point in time.

Note: time series and cross-sectional are terms that describe entire datasets, not individual variables.

Slide 22:

Data sources: where how and when

Knowing where a dataset came from is a critical first step in data analytics. Not all data is created equal.

Examples of data sources:

* Transactional: most data owned by businesses and organizations fall into this category
* Survey: data from surveys vary greatly in quality. Knowing the survey method is crucial. We’ll talk more about this in chapters 8 and 9. Random internet surveys are worthless
* Experiment: This can be the highest quality data available, but knowing the design of the experiment is critical.
* Publicly available data: Census, WHO, NYC open data

Example: St. John’s event data. The question was posed: how does event attendance correlate with student retention and persistence at St. John’s? We have tables that record event attendance, but the way attendance is taken is unclear. Not all events take attendance. Some take it using card swipes, and this is ideal. Others take it using handwritten sign-in sheets and data is entered manually – not ideal. How do we manage this? Unless we know how event attendance is taken, we can’t really get started.